DOCUMENT RESUME

ED 422 349 TM 028 901

AUTHOR Gyagenda, Ismail S.; Engelhard, George, Jr.
TITLE Rater, Domain, and Gender Influences on the Assessed Quality of Student Writing Using Weighted and Unweighted Scoring.
PUB DATE 1998-04-00
NOTE 33p.; Paper presented at the Annual Meeting of the American Educational Research Association (San Diego, CA, April 13-17, 1998).
PUB TYPE Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Essay Tests; Evaluators; *High School Students; High Schools; *Scoring; *Sex Differences; State Programs; Testing Programs; Writing (Composition); *Writing Tests
IDENTIFIERS Domain Knowledge; Georgia; *Rater Effects; *Weighting (Statistical)

ABSTRACT

The purpose of this study was to examine rater, domain, and gender influences on the assessed quality of student writing using weighted and unweighted scores. Twenty rates were randomly selected from a group of 87 operational raters contracted to rate essays as part of the 1993 field test of the Georgia High School Writing Test. All of the raters rated the complete set of 375 essays written by high school students. Each essay was scored on four domains and a statewide committee assigned the following judgmental weights (in parentheses) to each of the domains: content/organization (4); style (2); conventions (2); and sentence formation (2). The total scores and domain scores in the unweighted and weighted forms were dependent variables, while rater and gender were the independent variables. Results from the analysis of variance and multivariate analysis of variance analyses indicated significant rater and gender differences using both weighted and unweighted domain and total scores. The univariate rate/gender interaction effect was not significant, but the multivariate rater/gender effect was significant. (Contains 2 tables, 7 figures, and 34 references.) (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 422 349

RATER, DOMAIN, AND GENDER INFLUENCES ON THE ASSESSED QUALITY OF STUDENT WRITING USING WEIGHTED AND UNWEIGHTED SCORING

Ismail S. Gyagenda

and

George Engelhard, Jr.

Emory university

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

Ismail Gyagenda

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Address: Emory University
Division of Educational Studies
North Decatur Building
Atlanta, GA 30322

(404) 727-0622
(404) 251-9568 (H)

TM028901

Running head: Weighted scoring and writing assessment

Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA; April 13-17, 1998.

Abstract

The purpose of this study was to examine rater, domain, and gender influences on the assessed quality of student writing using weighted and unweighted scores. Twenty raters were randomly selected from a group of 87 operational raters contracted to rate essays as part of the 1993 field test of the Georgia High School Writing Test. All of the raters rated the complete set of three hundred seventy five essays written by high school students. Each essay was scored on four domains and a statewide committee assigned the following judgmental weights (in parentheses) to each of the domains: content/organization (4), style (2), conventions (2), and sentence formation (2). The total scores and the domain scores in the unweighted and weighted forms were the dependent variables, while rater and gender were the independent variables. Results from the ANOVA and MANOVA analyses indicated significant rater and gender differences using both weighted and unweighted domain and total scores. The univariate rater/gender interaction effect was not significant, but the multivariate rater/gender interaction effect was significant.

RATER, DOMAIN, AND GENDER INFLUENCES ON THE ASSESSED QUALITY OF STUDENT WRITING USING WEIGHTED AND UNWEIGHTED SCORING

Performance assessments are gaining acceptance for various measurement applications (Welch & Miller, 1995). One of the areas in which these assessments have been widely applied is in the measurement of student writing ability at both state and national levels (Afflerbach, 1985; Applebee, Langer, & Mullis, 1985; Linn, Baker, & Dunbar, 1991; Zwick, Donoghue, & Grimes, 1993). Given the high-stakes nature of many of these large-scale performance assessments (Welch & Miller, 1995), researchers have been concerned not only with various factors that may influence performance assessments, but with the unintended consequences of these assessments (Bond, 1995).

With regard to student writing ability, the major factors influencing its measurement may be related to: (a) the assessment process itself, such as rater bias, rater severity, or rating method (Huot, 1990; Lunz, Wright, & Linacre, 1990); (b) student characteristics like "gender, age, ethnicity, race, social class, or opportunity to learn" (Engelhard, 1992, p. 175); (c) factors linked to the writing task itself such as prompt or domain (Ruth & Murphy, 1988) or (d) a combination of the above factors. In the next four sections, selected research related to raters, domains, gender, and weighted scoring will be briefly reviewed.

Raters

The influence of the assessment process on student writing ability has received some research attention. Using the Rasch model to explore rater differences, Lunz,

Wright, and Linacre (1990), Engelhard (1994), and Du and Wright (1997) found significant differences in rater severity even after raters had received considerable training.

Engelhard (1994) examined rater differences in assessing essays. From the 1990 administration of the Eighth Grade Writing Test in Georgia, a sample of 264 randomly selected compositions assessed by 15 raters formed the data for Engelhard's (1994) study. Fifty-one percent of the compositions were written by females and forty-nine percent by males. Each composition was scored by two operational raters on five domains (content/organization, style, sentence-formation, usage, and mechanics). The ratings of the validity committee were used to anchor the calibrations of the 15 operational raters. Three facets (writing competence, rater severity, and domain difficulty) were utilized to calibrate the raters using the FACETS computer program. Results indicated significant differences between the raters, $\chi_2(15) = 170.7$, $p < .01$ with a high reliability of separation index ($R = .87$). Unless adjustments were made, the writing competence of the students judged by severe raters would be underestimated.

Du and Wright (1997), using data from the 1993 direct writing assessment of the Illinois State Goal Programs (IGAP), explored rater, rating scale, and writing task effects and used the many-faceted Rasch-model to adjust student measures based on those effects. The data for the study comprised 1734 randomly selected essays by 867 students in grades 6 (27%), 8 (24%), and 10 (49%), and scored by 89 trained raters. Results from the study indicated that there were significant differences in rater severity, and that the scales and topics were further sources of significant differences in student scores. Du and Wright concluded that to ensure objective measurement of student writing ability, rater,

rating scale, and task variations had to be identified and student scores had to be adjusted accordingly.

Domain

Domains refer to aspects or characteristics of essay quality that are analyzed and separately scored. Such aspects may include, for example, language mechanics, style, sentence formation, spelling, or content/organization. Research on domain influences seems to be scanty. However, Gabrielson, Gordon, and Engelhard (1995) in their examination of the effects of task choice on the writing quality of 11th grade students used as the dependent variables the following four domains in each essay: (a) content and organization, (b) style, (c) conventions, and (d) sentence formation. Female students scored significantly higher than the males on all domains, with the largest gender differences on the conventions domain, which referred to the appropriate usage of the mechanics of standard American English. White students performed significantly higher than Black students on all four domains, with the largest differences on the conventions domain, followed by sentence formation, the style, and the content and organization domains. These results tallied with earlier research by Engelhard, Gordon, Walker, and Gabrielson (1994).

Gender

In their seminal review on gender differences, Maccoby and Jacklin (1974) reported that from about the age 11, "girls have greater verbal ability than boys" (p. 351) and score higher than boys "on tasks involving both receptive and productive language ...(including) ...creative writing" (p. 351). Subsequent research indicates that girls perform better in assessed writing than boys (Applebee et al., 1990; Doolittle &

Welch,1989; Gabrielson, Gordon, & Engelhard, 1995; Hedges & Nowell, 1995; Hyde & Linn, 1988; Randhawa, 1991; Schick, DeMasi, & Green, 1992) and Engelhard, Gordon, Walker, & Gabrielson (1994) found gender influences to be related to the nature of the writing task.

In their study on the influence of writing tasks, gender, and race on the quality of student writing, Engelhard et al. (1994) sampled 170,899 Black and White 8th grade students who had taken a statewide-writing assessment in Georgia. They found that all the three independent variables had significant effects on the quality of writing. Girls performed better than boys. Gabrielson et al. (1995) also found that female students wrote essays of higher quality than male students.

Hedges and Nowell (1995) examined the magnitude of gender differences in mental abilities. Using six large-scale data sets, with nationally representative samples collected over a 32-year period, they explored gender differences in mean scores, variance of these scores, and in the number of individuals with extremely high and low scores. Results from the study indicated that females did slightly better in reading comprehension, perceptual speed, and associative memory; males did better in math and social sciences, and much better in vocational aptitude tests; and females substantially did better in writing. Hedges and Nowell (1995) concluded that in spite of the stability of gender differences over the 32-year period, the causes of these differences remain unresolved.

In another study, Randhawa (1991) looked at patterns of change in gender differences in the academic achievement of the 4th, 7th, and 10th graders in math and language. With a sample of 1300 students in one region in Canada in 1987 and 1989,

Randhawa found that females did better in language and that this advantage remained relatively stable between 1878-1985.

Doolittle and Welch (1989) examined gender differences in performance on the Collegiate Assessment of Academic Proficiency (CAAP). Students, mainly incoming college freshmen, were tested in reading, writing, mathematics, and critical thinking. The writing assessment included multiple-choice test and an essay test. Females performed significantly better than males on both writing tests.

A Conceptual Framework for Weighted Scoring

Weighting in writing assessment is manifested in different ways. It may be explicit judgmental, implicit judgmental, explicit empirical, or implicit empirical (see Fig. 1). In explicit judgmental weighting, test developers and educational administrators for various reasons often decide that certain domains should carry more weight than others. For example, in the Georgia High School graduation writing assessment, the Content/Organization domain carries twice as much weight as each of the other three domains: Style, Conventions, and Sentence Formation

Another type of explicit judgmental weighting that is commonly used by teachers and professors is to give weights to different types of assignments within one course or subject. The composite score would therefore have to be computed, taking into account the weights for each assignment. Ebel and Frisbie (1991) discuss appropriate ways of handling such weighting to come up with an objective and fair composite score for each student. They warn against merely summing up the final scores from each assignment, and recommend taking into account the standard deviation within each assignment category. The observed scores would be converted to T-scores, and these would be

multiplied by whatever weights are desired and summed to yield a total score on which grades would be assigned.

On the other hand, in implicit judgmental weighting, raters place extra value on certain aspects of the essay that may not be explicitly specified in the scoring rubric. Researchers have reported, for example, that essay raters tend to base their judgments primarily on the content and organization of the essay (Breland and Jones, 1984; Freedman, 1977). Raters also tend to give implicit weight to the form in which the essay is written (Chase, 1986; Powers et al., 1994; Sweedler-Brown, 1991). Sweedler-Brown (1991) examined the holistic scoring of handwritten and computer typed essays, and reported rater bias against typed final essays, especially those that had received high scores in the handwritten form. Powers et al. (1994) explored the impact on essay scores of intermingling handwritten and word-processed student essays. Each essay was converted to the other format and rescored. Handwritten essays received higher average scores than word-processed essays. It appears, therefore, that raters' implicit judgmental weights can influence examinee scores.

Another type of weighting is explicit empirical or statistical weighting (Wainer & Lukhele, 1997). Wainer and Lukhele (1997) suggest a method of item weighting that seeks to minimize the influence of an item's differential item functioning (DIF). DIF refers to the "extent to which there are idiosyncratic group differences in performance on an item" (Wainer & Lukhele, 1997, p. 201). The authors suggest, with a model example, a statistical method-differential DIF weighting-based on item response theory (IRT) "that weights each item by the size of its relation to the underlying trait being measured" (p. 203). They contend that this method, which is intended for large items (essays), weights

items by their "informational contribution rather than some a priori scheme" (p. 204) and would "reduce the effect that a differentially functioning large item has on total score without adversely affecting the precision of the test" (p. 205).

In the last category of implicit empirical weighting may fall everything else. All scoring may be conceived as involving weighting. Scoring involves assigning specific values to responses based on some pre-arranged explicit scheme or on some implicit notion of right, half-right, wrong or half-wrong etc. response. Translating this "notion" into specific values (scores) involves implicit operations that can be conceived as weighting.

Evidently, therefore, weighting is frequently used for various purposes in performance assessment. In spite of this prevalent and varied usage of weighting in performance assessments, there seems to be no research on the consequences of such implicit and explicit judgmental weighting, and how weighting may interact with other variables like raters, gender, and domain. There is a need to directly explore the influences of weighting on the assessed quality of student writing. This study focused on explicit judgmental weighting.

The purpose of this study, therefore, was to examine rater, domain, and gender influences on the assessed quality of student writing using both weighted and unweighted scores. The study addressed the following research questions:

1. Are there significant differences between raters using unweighted versus weighted total scores?
2. Are there significant gender differences in the assessed quality of student writing after controlling for the rater effect, using unweighted versus

weighted total scores?

3. Are there significant differences between the raters using unweighted versus weighted domain scores?
4. Are there significant gender differences in the assessed quality of student writing after controlling for the rater effect, using unweighted versus weighted domain scores?

METHOD

Participants

Twenty raters were randomly selected from a group of 87 operational raters contracted to rate essays as part of a 1993 field test of the Georgia High School Writing Test. Three hundred seventy five high school students participated in the study. Nine cases were dropped from the analyses because they contained missing values, leaving three hundred sixty six participants, with 197 male and 169 female students and their demographic characteristics were as follows: 46.7% female and 53.3% male; 77.7% White, 17.3% Black, and 5.0% Other.

Instrument

The Georgia High School Writing Test is intended to provide a direct assessment of student writing competence. Following the 1993 field test, this instrument has been used for high school graduation requirements in Georgia. Students were asked to write a composition of two pages (maximum) on an assigned prompt with a time limit of about 1 hour and 30 minutes. The student essays were analytically scored in the following four domains: content/organization, style, conventions, and sentence formation. The

content/organization domain measures student competence in the developing of a central idea. The style domain measures student use of language to establish individuality. The conventions domain focuses on student demonstrated ability to use the acceptable conventions of standard written English. The sentence formation domain measures student competence in writing correct sentences. A 4-point scale was used to obtain the unweighted scores on each of the domains. For a full description of the Georgia High School Writing Test, an assessment and instructional guide is available (Georgia Department of Education, 1993).

Procedures

Three hundred seventy five composition papers were selected by a statewide committee as benchmarks to represent the full range of scores available to the operational raters. The twenty operational raters scored the papers written by Georgia High School students. The statewide committee assigned the following judgmental weights (in parenthesis) to each of the domains: content/organization (4), style (2), conventions (2), and sentence formation (2). Because each domain was scored on a 4-point scale, their respective observed ratings had the same range from 1 to 4. Therefore, to obtain the weighted total score, the observed rating for each domain will be multiplied by its respective weight, and the products will be summed: Total Weight Score = (R1 x 4) + (R2 x 2) + (R3 x 2) + (R4 x 2), where R1...R4 are the four separate ratings in each domain. Here, the maximum possible range for the weighted scores is from 10 to 40. The total unweighted score is the sum of the observed ratings on each domain and its maximum range is from 4 to 16. The total scores and the domain scores in the unweighted and weighted forms will be used in the statistical analysis. To examine the

rater and gender influences on the unweighted and weighted total scores, two 2-way ANOVAs (rater x gender) were conducted using the General Linear Model (GLM). The dependent variables will be the unweighted and weighted total scores. To examine the rater and gender influences on the unweighted and weighted domain scores, two 2-way MANOVAS (rater x gender) will be conducted. The dependent variables will be the unweighted and weighted domain scores.

RESULTS

Regarding the total scores, there were significant rater differences (see Table 1) using both the unweighted ($F(19, 366) = 5.00$, $p < .001$) and the weighted total scores ($F(19,366) = 5.39$, $p < .001$). The overall rater means and standard deviations (in parentheses) for unweighted and weighted total scores were 10.83 (3.26) and 26.91 (8.17) respectively. The rater means ranged from 10.14 to 11.48 for unweighted scores, and 25.02 to 28.64 for the weighted scores.

There were also significant gender differences, after controlling for rater effects, with both unweighted ($F(1,366) = 897.13$, $p<.001$) and weighted total scores ($F(1,366) = 913.95$, $p<.001$). Female students performed better than males, with means (SDs in parentheses) on the unweighted total scores of 11.99 (2.72) and 9.84 (3.35) respectively.

Female students' means and standard deviations by rater on unweighted total scores ranged from 11.02 to 12.73 and 2.20 to 3.04 respectively. In comparison, means and standard deviations for male students ranged from 9.25 to 10.56 and 3.03 to 3.63 respectively. On the weighted total scores, the mean for females was 29.83, with a range

of 27.21 to 31.63, and the SD ranged between 5.74 and 7.63. The mean for males was 24.40, with a range of 22.85 to 26.38 and the SD ranged from 7.69 to 8.95.

However, the rater/gender interaction effect on both the unweighted and the weighted total scores was not significant.

With regard to domain scores (see Table 2), the multivariate analysis indicated significant rater differences using both the unweighted ($F(76, 366) = 16.34$, $p < .001$) and the weighted ($F(76, 366) = 16.34$, $p < .001$) domain scores. There was also a significant gender effect with both unweighted ($F(4, 366) = 236.73$, $p < .001$) and weighted ($F(4, 366) = 236.73$, $p < .001$) domain scores. The rater by gender interaction effect was also significant using the unweighted ($F(76, 366) = 1.42$, $p < .01$) and weighted ($F(76, 366) = 1.42$, $p < .01$) domain scores. It is interesting to note that the F values for these data are equivalent under the unweighted and weighted conditions.

The univariate analyses showed significant rater differences on all the four domains using both unweighted and weighted domain scores. The F (19, 366) values for the rater effect on content/organization, style, conventions, and sentence formation were 8.19, 10.14, 6.98, and 7.33 respectively in both the unweighted and weighted form. On content/organization, rater means, with standard deviations in parentheses, ranged from 2.36 (.92) to 2.84 (.91) and the overall rater mean was 2.63 (.91) using unweighted scores. On style, the rater means ranged from 2.33 (.88) to 2.84 (.93), with the overall rater mean of 2.54 (.91). On conventions, rater means ranged from 2.51 (.86) to 2.96 (.83) and the overall rater mean was 2.73 (.88). On sentence formation, the rater means ranged from 2.74 (.86) to 3.15 (.93), with the overall rater mean of 2.94 (.91).

The univariate gender differences were also significant at $p < .001$ in all the four domains using both unweighted and weighted scores, with female students scoring higher than boys in all domains. The $F(1, 366)$ values for the gender effect were 803.69 for the content/organization domain, 605.90 for style, 730.91 for conventions, and 756.50 for sentence formation. Using unweighted scores, female means and standard deviations (in parentheses) ranged from 2.58 (.71) to 3.19 (.86) on content/organization, from 2.58 (.65) to 3.20 (. 89) on style, from 2.79 (.61) to 3.22 (.88) on conventions, and from 2.99 (.67) to 3.50 (.85) on sentence formation. Similar statistics for male students ranged from 2.18 (.85) to 2.62 (.94) on content/organization, from 2.14 (.83) to 2.62 (1.00) on style, from 2.27 (.78) to 2.74 (1.01) on conventions, and from 2.49 (.85) to 2.85 (1.11) on sentence formation.

As earlier reported, the multivariate rater by gender interaction effect on the domain scores was significant, but the univariate rater by gender interaction effects were not significant. To examine this interaction, effective sizes (ES) related to all the dependent variables were calculated and plotted. These ESs clearly illustrate that some raters have larger gender effects than others. For example, in Fig. 2, even though the interaction effect is not statistically significant for total scores, the size of the gender differences range from an effect size of .39 for Rater 8 to an effect size of .62 for Raters 2 and 4 (these effect sizes are essentially the same with weighted scoring). Girls have higher scores for each rater, but the data also suggest that the size of this gender difference is a function of who rates the papers.

This rater by gender interaction is statistically significant when the domains are examined with a multivariate model. Figures 3 to 6 show the effect sizes by rater for

each domain: The effect sizes for the weighted scores were exactly the same and are not shown here. These figures clearly show that the size of the estimated gender differences varies by rater. For example, in the Content/Organization domain (Fig. 3), the effect sizes range from .32 for Rater 8 to .64 for Rater 2. In the Style domain (Fig. 4), the range of the effect sizes is from .28 for Rater 17 to .57 for Rater 2. In the Conventions domain (Fig. 5), the effect sizes range from .32 for Rater 15 to .56 for Rater 4. In the Sentence Formation domain (Fig. 6), the range of the effect sizes is from.37 for Rater 3 to .58 for Rater 4 . These figures also show that the pattern of rater by gender differences varies across the four domains. Figure 7, which combines all the domains, clearly illustrates that the raters have different, even contrasting effect sizes for the different domains. For example, the gender difference effect sizes for Rater 13 oscillate from .43 in Content/Organization domain to .33 in the Style domain, to .46 in the Conventions domain, to .56 in the Sentence Formation domain. Raters 20, 12, and 3 also exhibit large differences in their effect sizes for the different domains.

DISCUSSION

The purpose of this study was to explore the influences of rater, gender, domain, and weighted scoring on student writing assessment. The finding of significant rater differences means that, unfortunately, the scores that a boy or a girl gets on the essay may depend on which rater graded their composition. The intensive training these raters received apparently did not eliminate the rater effect. Other studies have reported this seemingly persistent rater effect (Du & Wright, 1997; Engelhard, 1994).

The finding of significant gender differences tallies with earlier research that found girls performing better than boys in writing (Gabrielson et al. 1995, Hedges & Nowell, 1995). Although the cause of superior female performance in writing is not known, it may be pertinent to suggest closer teacher attention to boy' writing skills, with a view to improving them. This is especially so given that females were superior than males on all domains. The persistent gender gap in writing favoring females may be welcome as a counter balance to the persistent male superior performance on standardized math and science tests! However, it may be worthwhile to probe whether there are classroom practices that may be discouraging boys from excelling in writing. Do teachers (and students) consider writing a female preserve? Are males less penalized, less scrutinized, or less challenged in writing activities than females? What do male students feel about their writing efficacy? These are some of the questions that may be worth probing to help us understand this persistent gender gap in writing ability.

With regard to weighted scoring, commonly used by test developers, there has been apparently no systematic and empirical appraisal of its potential influences on the measurement process. It is not known whether weights magnify, reduce, or distort the size and extent of other influences such as rater, gender, or race. Weights may lead to unintended consequences if their use introduces a biasing effect singularly or in conjunction with other variables. It is important that test developers be aware of such influences and consequences as they decide on the use of judgmental weights. In this study, however, the use of judgmental weights had no singular effect on students' assessment. The substantive results were essentially the same using both unweighted and weighted scoring. This may be because females were superior on all domains, including

the content/organization that was weighted differently than the other three. If males had performed better on some domains, and these domains were underweighted, it probably would have produced a significant weighting effect.

It also remains unclear why the significant rater by gender interaction effect on domain scores in the MANOVA analysis was not replicated in the ANOVA analyses. It may be because the four domains were highly inter-correlated. Further exploration of the rater and rater by gender interaction effects in essay assessment is needed. The results showing different effect sizes by rater (Figs. 2 to 7) seem to indicate a rater by gender interaction effect that warrants further exploration. Apparently some raters produce larger gender differences than others, and this effect seems to be influenced by the domain being rated. The Rasch-model that is appropriate in examining individual characteristics may offer better insights on the qualities of the raters, especially those that exhibit gender bias.

More research on weighting is also needed. It may be interesting to see what happens if females and males excel in different domains or subjects and these are weighted differently. Would the underweighting of domains in which males excel produce a biasing effect favoring females or vice versa? Would different weighting scales have any effect? Since weighting, as conceptualized in Fig. 1, is widespread, its potential effects on test fairness merits continued research attention.

Given the increasing interest in performance assessment in the form of essay tests, this study's replication of significant rater effects implies that essay test developers and administrators still need to grapple with the problem of rater bias. Since rater training seems to be failing to eliminate this rater effect, additional steps may be needed to

address this problem. Qualitative surveys of raters may provide helpful insights into the complex process of rating.

The ultimate goal in essay assessment, indeed in educational measurement in general, is to ensure that students are fairly assessed, and that their scores do not depend on extraneous characteristics (gender, race, ethnicity, etc.) or on who grades them. This study indicates significant rater and gender effects, and a significant multivariate rater/gender interaction effect. Therefore, the quest for minimizing these effects, especially rater effects, must continue.

REFERENCES

Afflerbach, P. (1985). *The statewide assessment of writing*. Princeton, NJ: Educational Testing Service.

Applebee, A.N., Langer, J.A., Jenkins, L.B., Mullis, I., & Foertsch, M.A. (1990). *Learning to write in our nation's schools: Instruction and achievement in 1988 at grade 4, 8, and 12*. Princeton, NJ: Educational Testing Service.

Applebee, A.N., Langer, J.A., & Mullis, I. (1985). *Writing: Trends across the decade, 1974-1984*. Princeton, NJ: Educational Testing Service.

Bond, L. (1995). Unintended consequences of performance assessments: Issues of bias and fairness. *Educational Measurement: Issues and Practice, 14*(3), 21-24.

Breland, H.M. & Jones, R.J. (1984). Perceptions of writing skills. *Written Communication, 1*, 101-109.

Chase, C.I. (1986). Essay test scoring: Interaction of relevant variables. *Journal of Educational Measurement, 23*(1), 33-41.

Doolittle, A., & Welch, C. (1989). *Gender differences in performance on a College-Level achievement test*. ERIC Reproduction Service. Document No. ED306 237.

Du, Y. & Wright, B.D. (1997). Measuring student writing abilities in a large-scale writing assessment. In M. Wilson, G. Engelhard, K. Draney (Eds.), *Objective Measurement: Theory Into Practice* (pp. 1-24). Norwood, NJ: Abex Publishing Corporation.

Ebel, R.L. & Frisbie, D.A. (1991). *Essentials of educational measurement* (5th ed.). Engelwood Cliffs, NJ: Prentice Hall.

Engelhard, G. (1992). The measurement of writing ability with a many-faceted Rasch model. *Applied Measurement in Education, 5*(3), 171-191.

Engelhard, G. (1994). Examining rater errors in the assessment of written composition with a many-faceted Rasch model. *Journal of Educational Measurement, 31*(2), 93-112.

Engelhard, G. (1996). Evaluating rater accuracy in performance assessments. *Journal of Educational Measurements, 33*(1), 56-70.

Engelhard, G., Gordon, B., Walker, E.V., & Gabrielson, S. (1994). Writing tasks and gender: influences on writing quality of black and white students. *Journal of Educational Research, 87*, 197-209.

Freedman, S.W. (1977). *Influences on the evaluators of student writing.* Dissertation Abstract International, 37, 5306A.

Gabrielson, S., Gordon, B., & Engelhard, G. (1995). The effects of task choice on the quality of writing obtained in a statewide assessment. *Applied Measurement in Education, 8*(4), 273-290.

Georgia Department of Education. (1993). *Georgia High School writing test: Assessment and instructional guide.* Atlanta, GA: Author.

Hedges, L.V. & Nowell, A. (1995). Sex differences in mental test scores, variability, and numbers of high-scoring individuals. *Science, 269*(5220), 41-45.

Huot, B. (1990). The literature of direct writing assessment: Major concerns and prevailing trends. *Review of Educational Research, 60*(2), 237-263.

Hyde, J.S. & Linn, M.C. (1988). Gender differences in verbal ability: A meta-analysis. *Psychological Bulletin, 104*, 53-69.

Klein, S.P., Jovanovic, J., Stecher, B.M., McCaffrey, D., Shavelson, R.J., Haertel, E., Solano-Flores, G., & Comfort, K. (1997). Gender and racial/ethnic differences on performance assessments in science. *Educational Evaluation and Policy Analysis, 19*(2), 83-97.

Linn, R.L., Baker, E.L., & Dunbar, S.B. (1991). Complex, performance-based assessment: Expectations and validation criteria. *Educational Researcher, 20*(8), 15-21.

Lunz, M.E., Wright, B.D., & Linacre, J.M. (1990). Measuring the impact of judge severity on examination scores. *Applied Measurement in Education, 3*, 331-345.

Oppler, S.H., Campbell, J.P., Pulakos, E.D., & Borman, W.C. (1992). Three approaches to the investigation of subgroup bias in performance measurement: review, results, and conclusions. *Journal of Applied Psychology, 77*, 201-217.

Popham, J. (1991). Interview on assessment items. *Educational Researcher, 20*(2), 24-27.

Powers, D.E. (1994). Will they think less of my handwritten essay if others word process theirs? Effects on essay scores of intermingling handwritten and word-processed essays. *Journal of Educational Measurement, 31*(3), 220-233.

Purves, A. & Hawisher, G. (1990). Writers, judges, and text models. In R. Beach & S. Hynds (Eds.), *Developing discourse practices in adolescence and adulthood* (pp. 183-199). Norwood, NJ: Ablex.

Randhawa, B. (1991). Gender differences in academic achievement: A closer look at Mathematics. *Alberta Journal of Educational Research, 37*(3), 241-257.

Ruth, L., & Murphy, S. (1988). *Designing writing tasks for the assessment of writing*. Norwood, NJ: Ablex.

Schick, R., DeMasi, M.E., & Green, M.S. (1992). Factors predicting writing performance. In A.C. Purves (Ed.), *The IEA study of written composition II: Education and performance in fourteen countries* (pp. 153-167). New York: Pergamon Press.

Sachse, P.P. (1984). Writing assessment in Texas: Practices and problems. *Educational Measurement: Issues and Practices, 3*, 21-23.

Sweedler-Brown, C.O. (1991). Computers and assessment: The effect of typing versus handwriting on the holistic scores of essays. *Research & Teaching in Developmental Education, 8*(1), 5-14.

Wainer, H. & Lukhele, R. (1997). Managing the influence of DIF from big items: The 1988 Advanced Placement History Test as an example. *Applied Measurement in Education, 10*(3), 201-215.

Welch, C. J. & Miller, T.R. (1995). Assessing differential item functioning in direct writing assessments: Problems and an example. *Journal of Educational Measurement, 32*(2), 163-178.

Zwick, R., Donoghue, J.R., & Grimes, A. (1993). Assessment of differential item functioning for performance tasks. *Journal of Educational Measurement, 30*(3), 233-251.

Table 1

Summary of Analysis of Variance on Unweighted and Weighted Total Scores

	Unweighted			Weighted		
Source	df	SS	F value	df	SS	F value
Rater	19	890.95	5.00*	19	6022.08	5.39*
Gender	1	8412.73	897.13*	1	53715.25	913.95*
Rater*Gender	19	186.72	1.05	19	1190.85	1.95
Error	7280	6826.03		7280	427863.81	
Corrected Total	7319	77738.54		7319	488691.40	
R-Square	0.12			0.12		

Note. F values are based on Type III sequential SS.
p< .001

Table 2

Summary of Multivariate and Univariate Analyses for Unweighted and Weighted Domain Scores

		Unweighted Domains					Weighted Domains				
Source		Multivariate Analysis	Univariate Analyses				Multivariate Analysis	Univariate Analyses			
			Content/Organization	Style	Conventions	Sentence Formation		Content/Organization	Style	Conventions	Sentence Formation
Rater	df	76	19	19	19	19	76	19	19	19	19
	SS		112.97	143.02	92.79	103.71		1807.51	572.86	371.16	414.82
	F	16.34*	8.19*	10.14*	6.98*	7.33*	16.34*	8.19*	10.14*	6.98*	7.33*
Gender	df	4	1	1	1	1	4	1	1	1	1
	SS		583.79	449.98	551.40	563.21		9340.68	1799.92	2045.63	2252.85
	F	236.73*	803.69*	605.90*	730.91*	756.50*	236.73*	803.69*	605.90*	730.91*	756.50*
Rater*Gender	df	76	19	19	19	19	76	19	19	19	19
	SS		16.22	18.57	14.65	15.08		259.55	74.29	58.59	60.33
	F	1.42*	1.18	1.32	1.10	1.07	1.42*	1.18	1.32	1.10	1.07
Error	df		7280	7280	7280	7280		7280	7280	7280	7280
	SS		5288.15	5406.62	5093.73	5419.95		84610.34	21626.48	20374.90	21679.81
Corrected Total	df		7319	7319	7319	7319		7319	7319	7319	7319
	SS		6001.12	6016.12	5712.72	6101.21		96017.99	24064.46	22850.87	24404.85

Note. F values are based on Type III sequential SS. (Results of the univariate and multivariate F tests are identical under the unweighted and weighted conditions).
*P< .01

Fig. 1

A Conceptual Framework for Weighted Scoring

	Judgmental	Empirical
Explicit	1. Educators, expert panels, or test developers assign weights to items or domains (e.g. the Georgia data used in this study). 2. Teachers or professors give weights to different assignments within a course.	1. Weights derived from various statistical methods e.g. 2-parameter IRT or item discrimination (Wainer & Lukhele, 1997). 2. All regression analyses 3. Statistical weighting (Ebel & Frisbie, 1991)
Implicit	Raters assign weights that are not specified in the scoring rubric e.g. handwritten versus typed essays (Sweedler-Brown, 1991).	Everything else. All scoring involves weighting of some sort

Figure 2
Gender differences (effect size) by rater for unweighted and weighted total scores

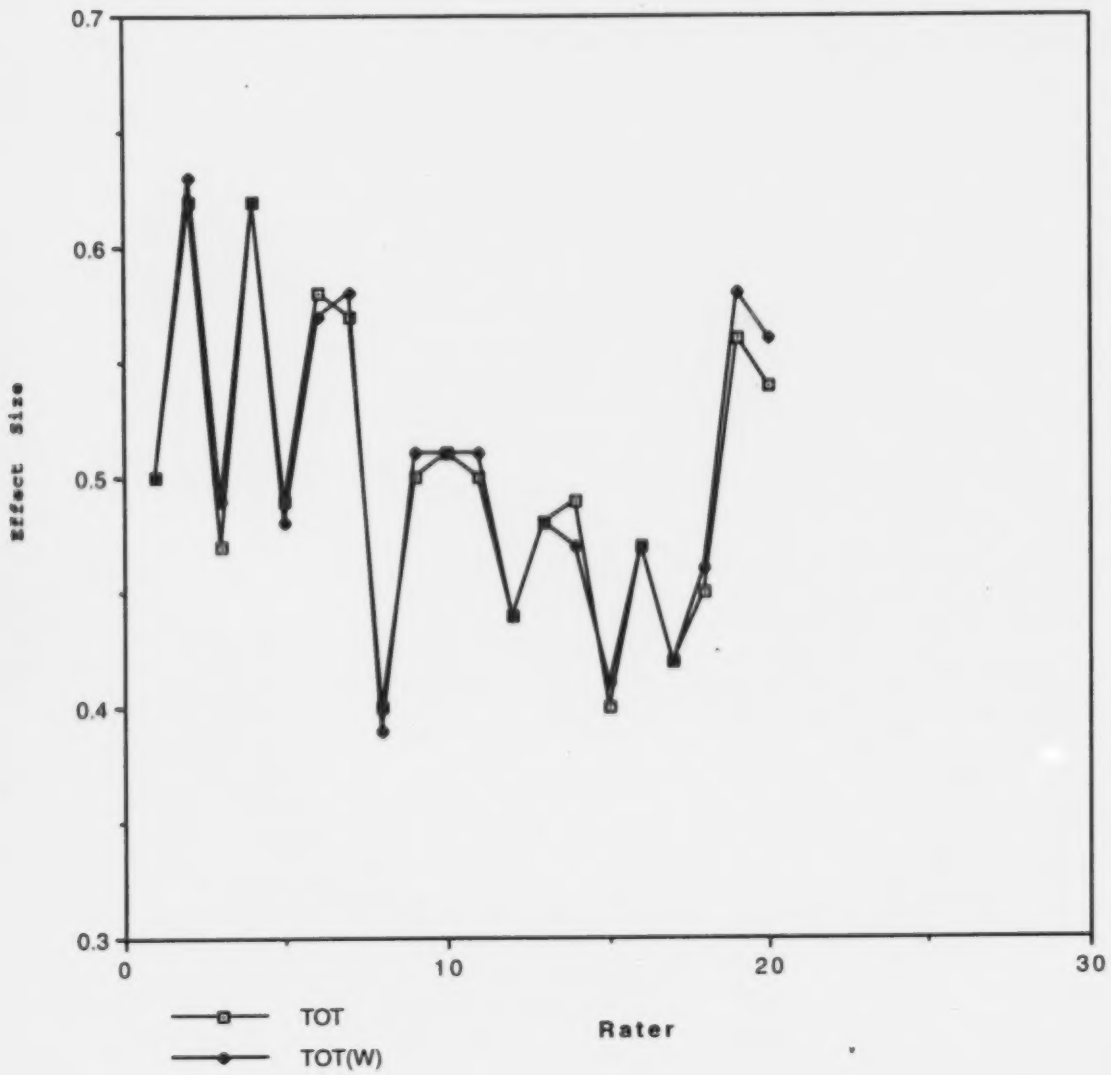

Figure 3
Gender differences (effect sizes) by rater for Content/Organization

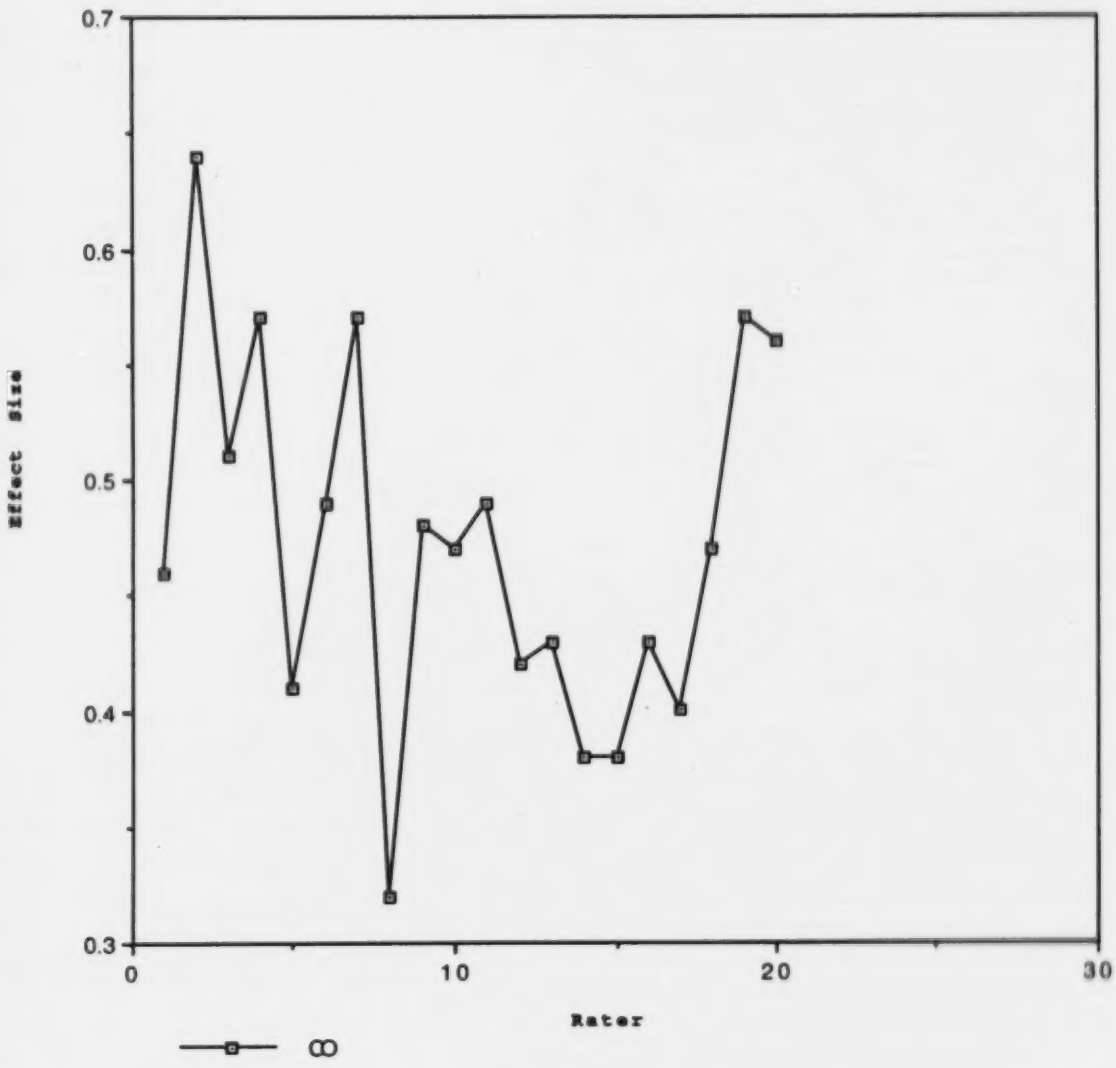

Figure 4
Gender differences (effect size) by rater for Style

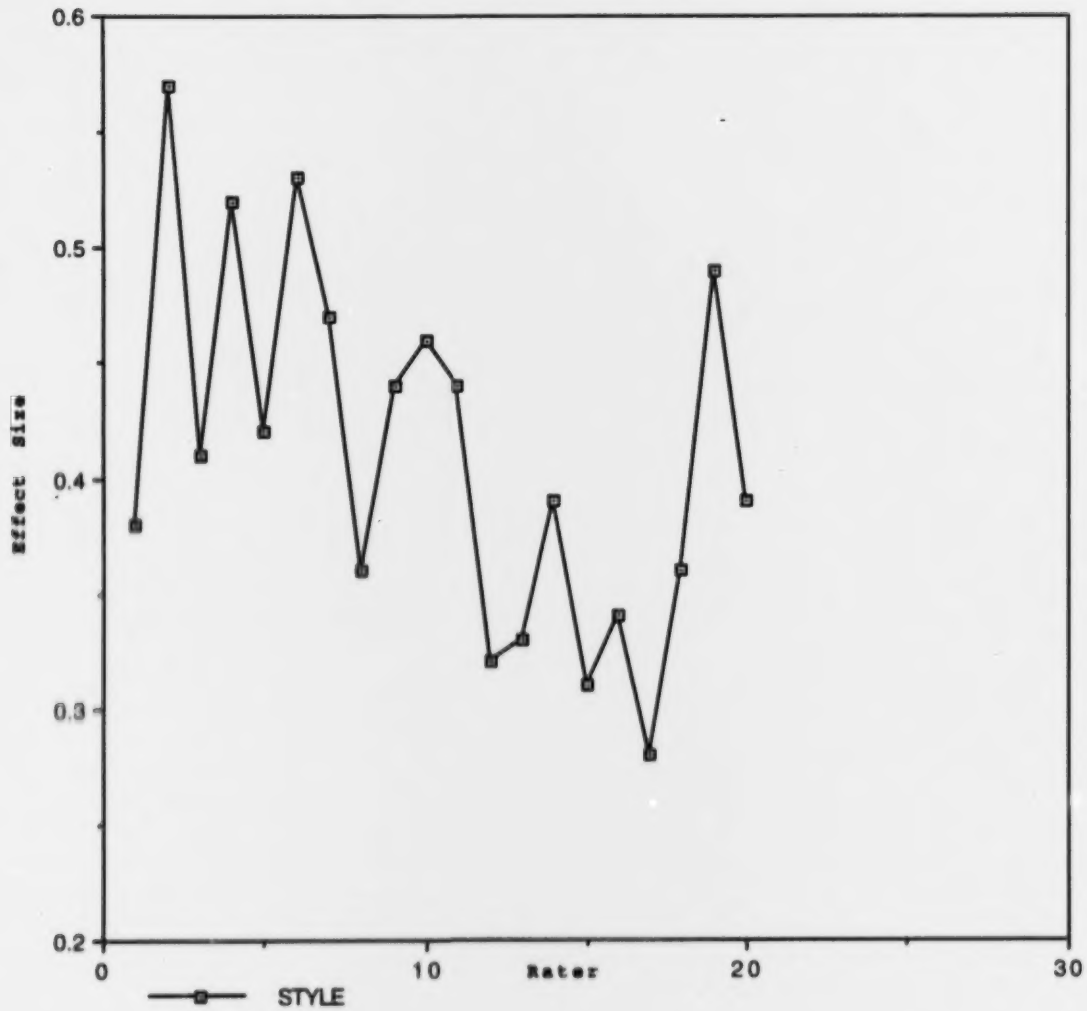

Figure 5
Gender differences (effect size) by rater for Conventions

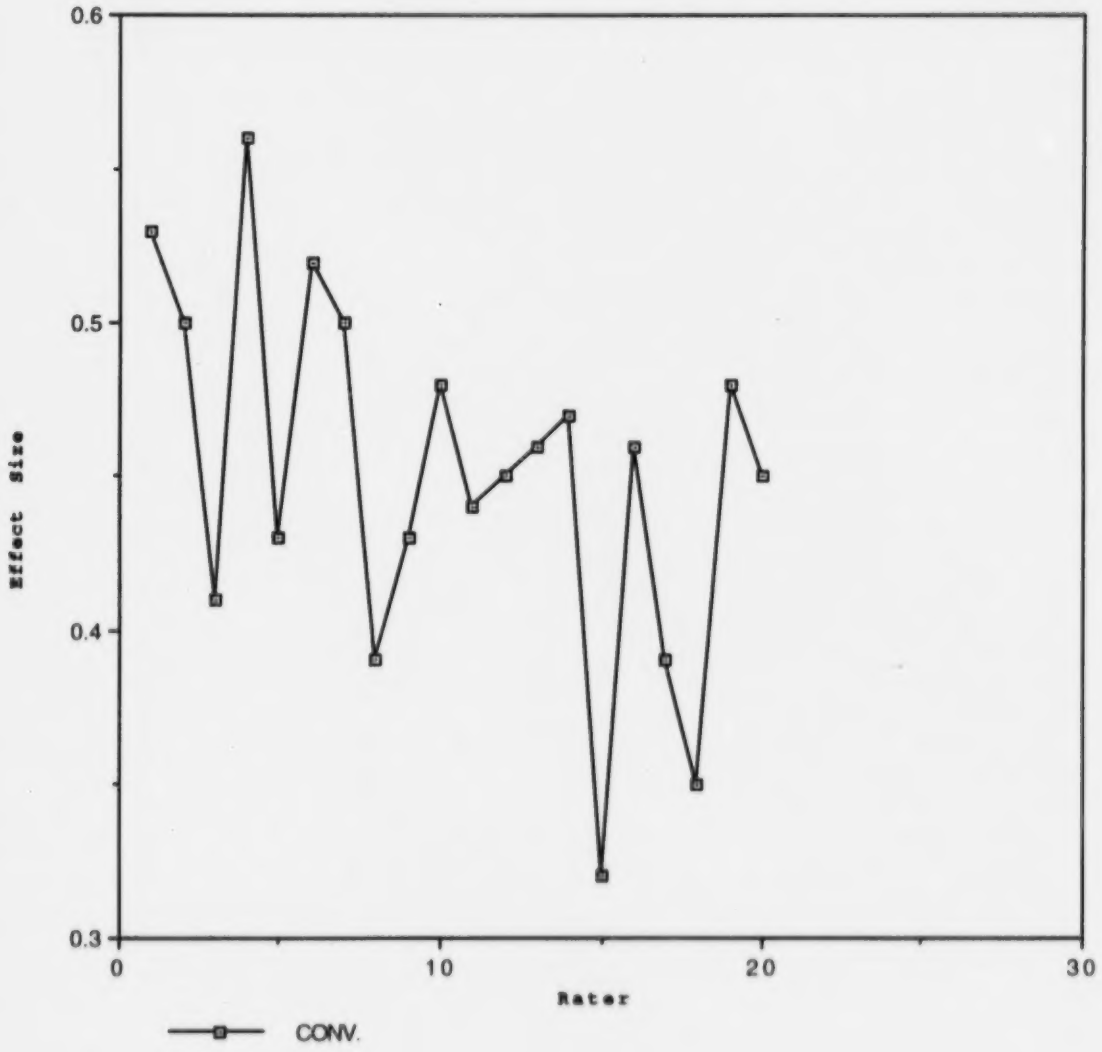

Figure 6
Gender differences (effect size) by rater for Sentence Formation

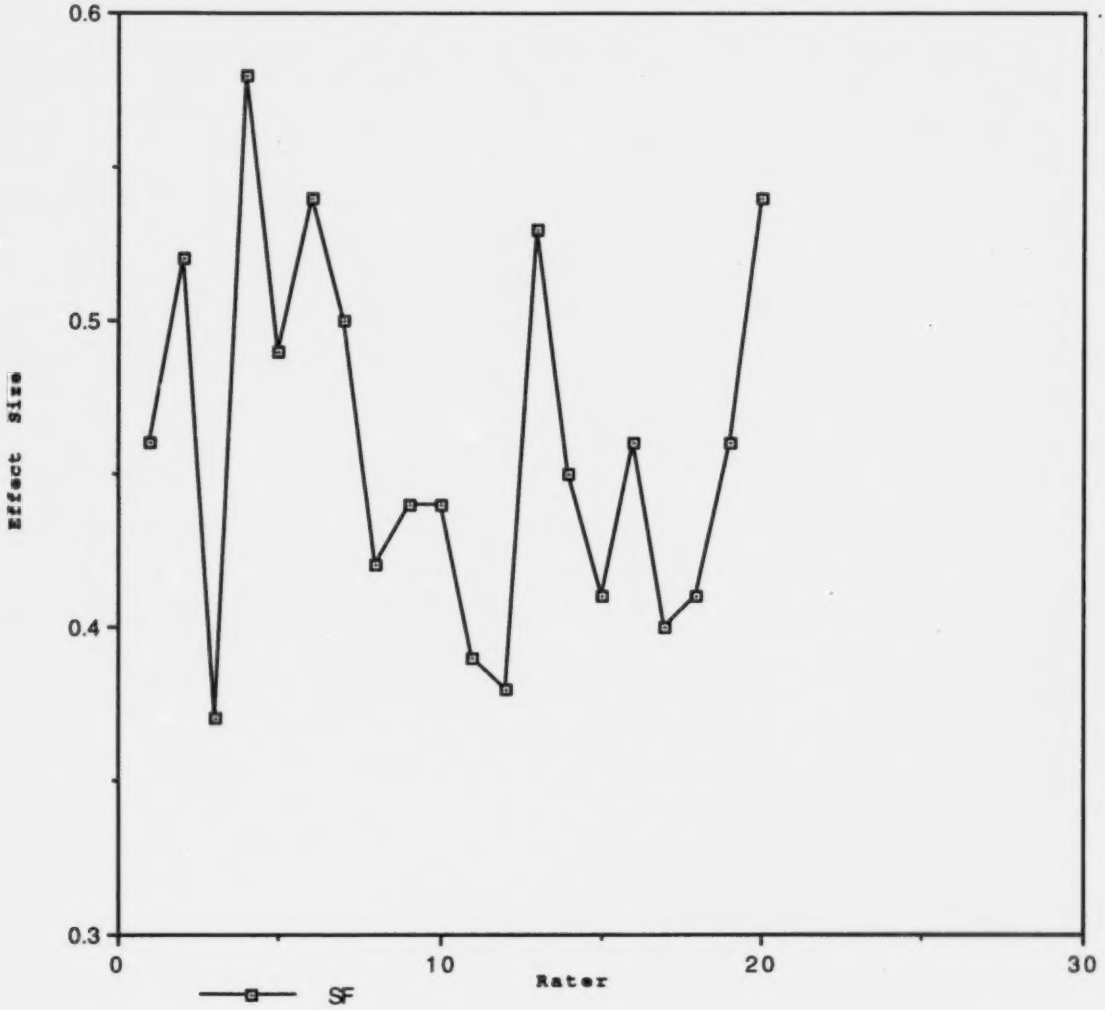

Fig. 7
Gender differences (effect size) by rater for all domains